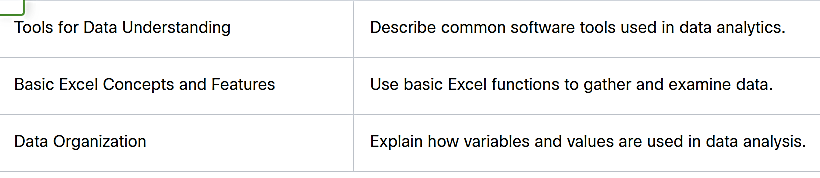
**Data Gathering and Investigation**

**2.0.1 Introduction**

In this module, you will move to the next two steps of the Data Analysis Process, **Obtaining and Investigating Data**. You will be **introduced to Kaggle,** a popular online community where data analysts, scientists and AI enthusiasts collaborate. You will also get some valuable **hands-on experience using Excel,** one of the world’s most used spreadsheet technologies.



**2.1 Tools for Data Understanding**

**2.1.1 Analysis Overview**

**2.1.2 Data and Datasets**

Data is now being collected and shared across many different organizations and in many different formats; **a collection of data is referred to as a dataset**. Datasets **may exist for the private** use of an individual organization **or shared across** the internet to anyone who wants to reference them. An **example of a private dataset** is a **physician’s patient dataset**, which might include patient demographics, test results, diagnosis, and appointment schedules. Access to this dataset is limited to those with permission to use it. In contrast, an **example of a publicly available dataset** is the World Health Organization **(WHO) open data repository**, which contains health-related statistics for its **194 member countries** and can be downloaded by anyone.

**Datasets often contain multiple related files stored in different formats**. Information about a dataset, including a **description of what it contains and how it is formatted, is called metadata.** Metadata files are valuable tools to provide analysts with an understanding of the data within the dataset.

One of the **most common formats** used to package and exchange data is the **Comma Separated Values (CSV) format**. Often, datasets that are publicly available may be made up of multiple CSV files that contain related data. These CSV files can be imported into tools such as Excel for further investigation and analysis. Later in this module, you will obtain and work with a CSV.

**2.1.3 Tools for Data Analysis**

In this course, you will have the opportunity to use three essential data analytics tools:

* Excel,
* SQL,
* and Tableau.

**MS-Excel**

What is Excel?

**Excel is a powerful tool** suitable **for small datasets** and quick data analysis. With Excel, you can **manipulate data, summarize it with pivot tables, visualize it, and perform quick statistics to summarize it.**

Why it’s important to know:

Excel is powerful and very popular for performing small-scale data analysis, calculations, data summaries, and data visualizations.

Excel skills you will learn in this course:

* Perform **data cleaning** by removing blank spaces as well as incorrect and outdated information
* Format and adjust data using **conditional formatting**
* Perform **data calculations** using **formulas**
* Organize data using **sorting and filtering**
* Create **visualizations** using graphing and charting
* **Calculate, summarize, and analyze data using pivot tables**
* **Aggregate** data for analysis

**SQL**

What is SQL?

SQL, which stands for **Structured Query Language**, is a powerful **database management tool** that allows data analysts to **retrieve and interact** with selections of **data that are stored in relational databases.** Relational databases have a defined structure and contain multiple interrelated data tables that need to be queried with a language like SQL to be useful. **SQL is fast** and can **handle data sets much larger than Excel can**. As a data analyst you will use **SQL to access, read, manipulate, and analyze** the data stored in a relational database to generate useful insights to drive a data-informed decision-making process.

Why it’s important to know:

Popular **big data systems make use of SQL** for maintaining relational databases and processing structured data. It is used for carrying out data analytics with data stored in relational database management systems such as Oracle, Microsoft SQL, and MySQL.

SQL skills you will learn in this course:

* **Create tables**
* Retrieve data using **SQL index**
* Retrieve data using **SQL queries**
* Aggregate data with **SQL joins**

**Tableau**

What is Tableau?

Tableau is one of the **most used data analytics and visualization tools** on the market. Visualizations are an important way to present data in a format that can easily be **understood by non-technical decision makers and stakeholders.**

Why it’s important to know:

* Tableau is a data analytics **market leader** due to the depth and quality of its data visualizations.
* **Tableau can extract** and combine data **from multiple sources** including **Excel** spreadsheets and **SQL** databases. It can also access large data storage locations, known as data warehouses, as well as cloud-based data repositories.

Tableau skills you will learn in this course:

* Compare data from multiple views using **Tableau dashboards**
* Create visualizations using **Tableau visualization tools**

**2.1.4 Introduction to Kaggle and Datasets**

**2.1.5 Lab - Introduction to Microsoft Excel**

Video.

**2.2.2 Lab - Basic Excel Concepts**

Lab.

**2.2.4 Lab - Importing Data into Excel**

Lab.

**2.2.6 Lab - Excel Practice**

Lab.

**2.3 Use simple functions for data analysis**

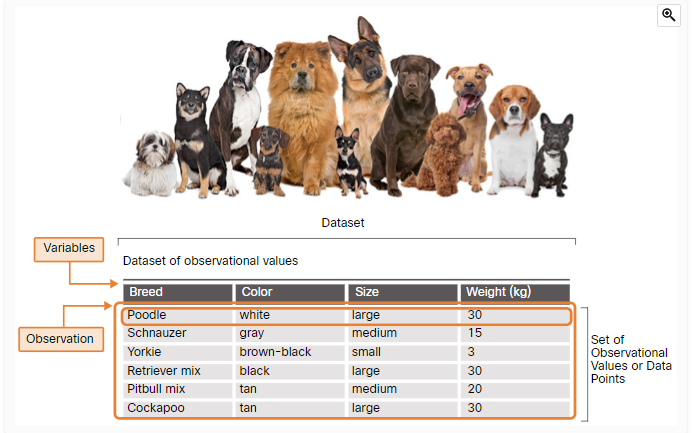
**2.3.2 Observations, Variables, and Values**

When performing any kind of data experiment or analysis, it is critical to define the key characteristics that need to be measured or observed. These characteristics to be studied are called variables. A **variable** in this context is anything that varies from one instance to another, that can be measured, and whose value can be manipulated or controlled in theoretical scenarios.

The recordings of the values, patterns, and occurrences for a set of variables are **observations**. The value or set of values for a specific observation is called a **data point**. Each observation can be thought of and represented as a record in a database or a row in an Excel spreadsheet. The collection of observations makes up the dataset for your analysis.

Observations usually have a purpose, and the variables included will depend on their relevance to that purpose. For example, if you have lost your pet and have asked other people to help you search for it, only a small set of variables—the dog’s characteristics—are relevant to their observations. These characteristics might be:

* **What type of animal is your pet?** It is a dog.
* **What type of dog?** It is a Schnauzer.
* **What color is your Schnauzer?**It is gray.
* **What size is the Schnauzer?** It is a medium sized Schnauzer.
* **How much does the Schnauzer weigh?** It weighs 15 kg.



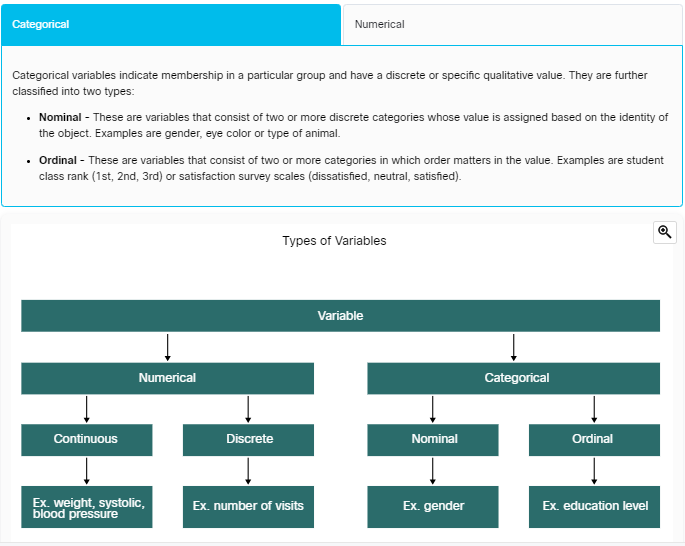
As shown in the figure, the variables are the characteristics, such as breed, color, size and weight. All of these characteristics are variables, because each can have multiple values. As people search for your dog, data points are added for each observation. Because the purpose of your observations is to search for your lost dog, observations that do not meet the required criteria are discarded.

**2.3.4 Types of Variables**

When looking for meaningful patterns in data, we frequently look for relationships between variables, as shown in the figure. All variables can be classified by the characteristic that is being studied.

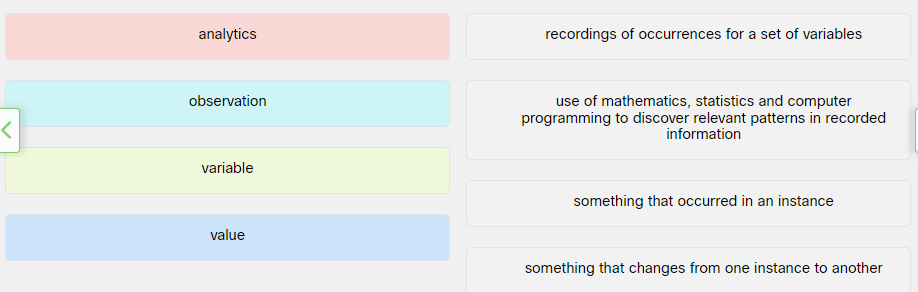
The variables will either be categorical or numerical.

**Select each variable type to learn more.**



Why is it important to know what types of variables are in your data set?

Some types of analysis and data visualizations are designed to work with certain types of data. How you might choose to present the results of the analysis will depend on the type of variables used in the data. Some types of variables lend themselves better to bar graphs, while others may allow for more examination and discovery using a scatter plot. Examples of some of the suggested types of graphs that represent the different types of variables can be seen in the figure above.



**2.3.4 Types of Variables**

When looking for meaningful patterns in data, we frequently look for relationships between variables, as shown in the figure. All variables can be classified by the characteristic that is being studied.

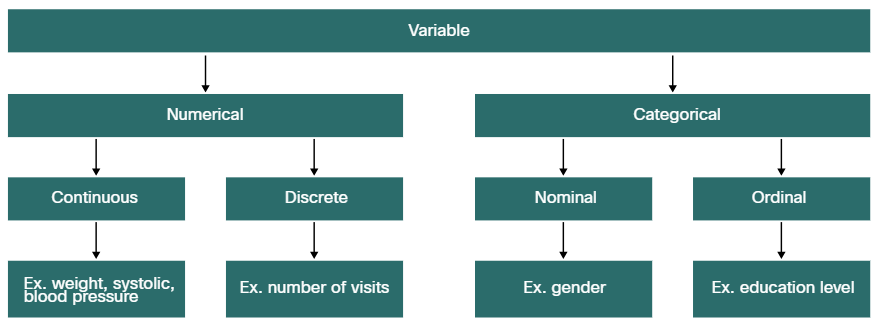
The variables will either be categorical or numerical.

Categorical variables indicate membership in a particular group and have a discrete or specific qualitative value. They are further classified into two types:

* **Nominal**- These are variables that consist of two or more discrete categories whose value is assigned based on the identity of the object. Examples are gender, eye color or type of animal.
* **Ordinal**- These are variables that consist of two or more categories in which order matters in the value. Examples are student class rank (1st, 2nd, 3rd) or satisfaction survey scales (dissatisfied, neutral, satisfied).

Numerical variables are quantitative values:

* **Continuous** - These are variables that are quantitative and can be measured along a continuum or range of values. There are two types of continuous variables: Interval variables can have any value within the range of values, and examples are temperature or time; Ratio variables are special interval variables where a value of zero (0) can mean that there is none of that variable and examples are income or sales volume.
* **Discrete**- These types of continuous variables are quantitative but have a specific value from a finite set of values. Examples include the number of sensors activated in a network, or the number of cars in a lot.



Why is it important to know what types of variables are in your data set? Some types of analysis and data visualizations are designed to work with certain types of data. How you might choose to present the results of the analysis will depend on the type of variables used in the data. Some types of variables lend themselves better to bar graphs, while others may allow for more examination and discovery using a scatter plot. Examples of some of the suggested types of graphs that represent the different types of variables can be seen in the figure above.

